课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	课程名称 Course Title	**数据挖掘 Data Mining**
2.	授课院系 Originating Department	计算机科学与工程系 Department of Computer Science and Engineering
3.	课程编号 Course Code	CS306
4.	课程学分 Credit Value	3
5.	课程类别 Course Type	专业选修课 Major Elective Courses
6.	授课学期 Semester	春季 Spring
7.	授课语言 Teaching Language	英文 English
8.	授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）	骆宗伟，副教授，计算机科学与工程系，luozw@sustech.edu.cn Zongwei Luo, Associate Professor, Department of Computer Science and Engineering, luozw@sustech.edu.cn
9.	实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact	
10.	选课人数限额(可不填) Maximum Enrolment （Optional）	

11. 授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
学时数 Credit Hours	32		32		64

12.	**先修课程、其它学习要求 Pre-requisites or Other Academic Requirements**	CS203 数据结构与算法分析 Data Structures and Algorithm Analysis 或 or CS203B 数据结构与算法分析 B Data Structures and Algorithm Analysis B
13.	**后续课程、其它学习规划 Courses for which this course is a pre-requisite**	无 None.
14.	**其它要求修读本课程的学系 Cross-listing Dept.**	无 Not applicable for other departments beside CS.

教学大纲及教学日历 SYLLABUS

15. **教学目标 Course Objectives**

1. 在数据挖掘的高级概念、算法和技术及其在大规模数据集和大数据分析中的应用方面获得广泛的知识。

2. 了解大数据挖掘和分析中的研究问题和主题。

1. Obtain broad knowledge in advanced concepts, algorithms and techniques for data mining and their applications to large-scale data set and big data analytics.

2. Understand research issues and topics in big data mining and analytics.

16. **预达学习成果 Learning Outcomes**

本课程的学生将能够将概念、算法和技术应用于大规模数据集和大数据分析。

1. 学生将具备准备海量数据源（>10G）的技能，并准备好用于数据挖掘。
2. 熟悉大型数据集挖掘的典型算法。
3. 熟悉大型数据集挖掘的典型应用。
4. 学生将能够选择使用大量数据挖掘技术解决计算问题的算法和/或方法。
5. 学生将能够设计新的算法和/或挖掘海量数据集的方法。

Students taking this course will be able to apply the concept, algorithms, and techniques in large scale data set and big data analytics.

1. Student will have the skills to prepare massive data source (>10G) and make it ready for use in data mining.
2. Student will be familiar with typical algorithms for mining large scale data set.
3. Student will be familiar with typical application of mining large scale data set.
4. Student will be able to select algorithms and/or methods in solving computing problems with massive data mining techniques.
5. Students will be able to design new algorithms and/or methods for mining massive data set.

17. **课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）**
Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

本课程提供数据挖掘的先进概念、算法和技术，以及它们在大规模数据集和大数据分析中的应用。介绍了数据处理/清理/分析、分类、关联分析、聚类分析和异常检测的基本算法，选择了数据挖掘和应用中的高级研究课题，重点介绍了时间数据、序列数据、空间数据、轨迹数据、图形数据、特克斯等各种数据类型。课程将涵盖实际数据、社会数据和各种应用程序。目标是为每个学生提供必要的技能，以便在数据分析领域进行进一步的学习和研究。（理论学时：32 实验学时：32）

第 1 周：导论：介绍数据挖掘，为课程做必要的准备。实验室将分配大型数据集的准备工作。

第 2 周：地图减少和软件堆栈：回顾大数据处理和相关软件架构。实验室的工作将是让学生熟悉软件工具。

第 3 周：查找相似项：将回顾选定的经典项相似性计算算法。实验室的工作将是让学生熟悉这些算法的实现和应用。

第 4 周：数据流挖掘：将介绍数据流挖掘算法。实验室的工作将是让学生熟悉这些算法的实现和应用。

第 5 周：经常项目集：将回顾识别经常项目集的挑战和重要性。实验室的工作将是让学生熟悉这些算法的实现和应用。

第 6 周：集群：将介绍集群的技术、应用程序和趋势，更多是在大数据分析的背景下。实验室的工作是让学生熟悉这些技术。

第 7 周：随机森林：将对随机森林的变量重要性、变量选择和异常值检测给予重要性。实验室工作将对学生进行课堂上所学技术的培训。

第 8 周：期中考试：学生成绩回顾。计划采用实验室形式。

第 9 周：社交网络：将介绍社交网络。实验室工作将集中在社交网络中的关键算法上。研究论文将分发给学生阅读。

第 10 周：病毒式营销：重点将放在通过协作数据挖掘的病毒式营销上，尤其是在社交网络上。将分发研究论文，进行实验室工作，让学生熟悉这些采矿技术。

第 11 周：推荐系统：将介绍协作过滤等关键推荐算法。研究论文将被分发，实验室工作将在论文中实施这些技术。

第 12 周：CNN 神经网络：将对 CNN 网络进行介绍。重点是 CNN 网络的原理理解和应用。将分发经典研究论文。实验室工作将加强这些主题。

第 13 周：RNN 神经网络：将介绍 RNN 网络。重点是对 RNN 网络的原理理解和应用。将分发经典研究论文。实验室工作将加强这些主题。

第 14 周：财务时间序列：将介绍财务数据处理。回顾传统的时间序列分析统计工具。实验室工作将在这些时间序列分析工具上进行。

第 15 周：财务时间序列：将介绍财务时间序列分析的关键绩效指标。将分发有关财务时间序列分析技术的研究论文。实验室工作将集中在选定的技术实施上。

第 16 周：学期和项目评审：对团队项目进行评审，并对整个课程进行评审。

This course provides advanced concepts, algorithms, and techniques for data mining and their application to large-scale data set and big data analytics. With introduction of fundamental algorithms for data processing/cleaning/analysis, classification, association analysis, cluster analysis, and anomaly detection, selected advanced research topics in data mining and applications, with emphasis on various data types such as temporal data, sequence data, spatial data, trajectory data, graph data, textual data, social data, and various applications, will be covered in the course. The goal is to provide every student necessary skill set to pursue further study and research in the data analytics field. （Theory Hours：32　Lab Hours：32）

Week 1: Introduction：an introduction to data mining, necessary preparation for the course. Preparation of large scale of

data sets will be assigned in the lab. .

Week 2: Map-reduce and software stack: review big data processing and related software architecture. Lab works will be on letting student familiar with software tools.

Week 3: Find similar items: selected classic item similarity computing algorithms will be reviewed. Lab works will be on letting student familiar with implementation and applications of those algorithms.

Week 4: Data steaming mining: data streaming mining algorithms will be introduced. Lab works will be on letting student familiar with implementation and applications of those algorithms.

Week 5: Frequent item set: challenges and importance of identifying frequent itemset will be reviewed. Lab works will be on letting student familiar with implementation and applications of these algorithms.

Week 6: Clustering: Techniques, Applications and Trends of clustering will be introduced, more in the context of big data analytics. Lab work will be on letting student familiar with these techniques.

Week 7: Random forests: Importance will be given on variable importance, variable selection, and outlier detection for random forests. Lab work will train students on those techniques learned in the classroom.

Week 8: Mid-term exam: Performance review of students. It is planned to be in the form of lab.

Week 9: Social networks: An introduction to social networks will be conducted. Lab work will be focused around key algorithms in social networks. Research papers will be distributed for students to read.

Week 10: Viral marketing: Focus will be on viral marketing via collaborative data mining, especially on top of social networks. Research papers will be distributed and lab work will be conducted to let student familiar with those mining techniques.

Week 11: Recommender systems: Key recommendation algorithms like collaborative filtering will be introduced. Research papers will be distributed and lab work will be around to implementing those techniques in the paper.

Week 12: CNN Neural networks: An introduction to CNN networks will be conducted. Focus will be on principle understanding and application of CNN networks. Classic research papers will be distributed. Lab work will strengthen those topics.

Week 13: RNN Neural networks: An introduction to RNN networks will be conducted. Focus will be on principle understanding and application of RNN networks. Classic research papers will be distributed. Lab work will strengthen those topics.

Week 14: Financial time series: An introduction to financial data processing will be conducted. Traditional statistical tools for time series analysis will be reviewed. Lab work will be on those time series analysis tools.

Week 15: Financial time series: Key performance indicators of financial time series analysis will be introduced. Research papers on financial time series analysis techniques will be distributed. Lab work will be focused on selected techniques implementation.

Week 16: Term and project review: Review of team project and review of the whole course will be conducted.

18. **教材及其它参考资料 Textbook and Supplementary Readings**

1. Jure Leskovec, Anand Rajaraman, Jeff Ullman Stanford University, Mining of Massive Datasets, Cambridge Press
2. David Easley, Jon Kleinberg, Networks, Crowds, and Markets – Reasoning about a Highly Connected World, Cambridge Press.

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance				
课堂表现 Class Performance				
小测验 Quiz		10%		随堂测验 In class quiz
课程项目 Projects		30%		小组工程 Group project
平时作业 Assignments		30%		试验作业 Lab assignment
期中考试 Mid-Term Test				
期末考试 Final Exam		30%		闭卷考试 Unseen exam
期末报告 Final Presentation				
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)				

20. 记分方式 GRADING SYSTEM

☑ A. 十三级等级制 Letter Grading
☐ B. 二级记分制（通过/不通过） Pass/Fail Grading

课程审批 REVIEW AND APPROVAL

21. 本课程设置已经过以下责任人/委员会审议通过
This Course has been approved by the following person or committee of authority